TASHKILOTLARDA HUJJATLARNI KLASTERLASHNING AVTOMATLASHTIRISH MODELI VA ALGORITMLARI

Axmedova Xolisxon Ilxomovna,
t.f.f.d., PhD
xolisa9029@mail.ru
O‘zbekiston xalqaro islomshunoslik akademiyasi

Mamadaliyeva Xadicha Baxodirjon qizi
4-kurs talabasi
xadichamamadaliyeva3@gmail.com
O‘zbekiston xalqaro islomshunoslik akademiyasi

Annotatsiya. Ushbu maqolada elektron hujjatlarni klasterlash va vektorlashning nazariy asoslari tahlil qilinadi. Matnli ma’lumotlarni qayta ishlash usullari hamda klasterlash algoritmlarining o‘ziga xos jihatlari yoritiladi. Tashkilotda hujjat almashish tizimini avtomatlashtirishni amalga oshirishda mashinali o‘qitish algoritmlaridan foydalanish mumkin. Ma’lumki, hozirda davlat tashkilotlari orasida hujjat almashish tizimi ishlab chiqilgan, ammo bir tashkilot ichida hujjatlarni bo‘limlarga taqsimlash, ya’ni devonxona bo‘limining vazifasini avtomatlashtirishni uchratmadik. Ushbu maqolada devonxona bo‘limiga kelib tushgan hujjatlarni tashkilotning mos kalsterlariga taqdim etish tizimini ishlab chiqish haqida so‘z boradi.

Kalit so‘zlar: *elektron hujjatlar, hujjatlarni klasterlash, matnni vektorlash, klasterlash algoritmlari, tabiiy tilni qayta ishlash, matn o‘xshashligi, K-means.*

Abstract: This article analyzes the theoretical foundations of clustering and vectorizing electronic documents. It highlights text data processing methods and the specific aspects of clustering algorithms. Machine learning algorithms can be utilized to automate the document management system within an organization. It is well known that document exchange systems have already been developed between government agencies; however, we have not encountered automation for distributing documents among departments within a single organization–specifically, the automation of the chancellery (records office) department's tasks. This article

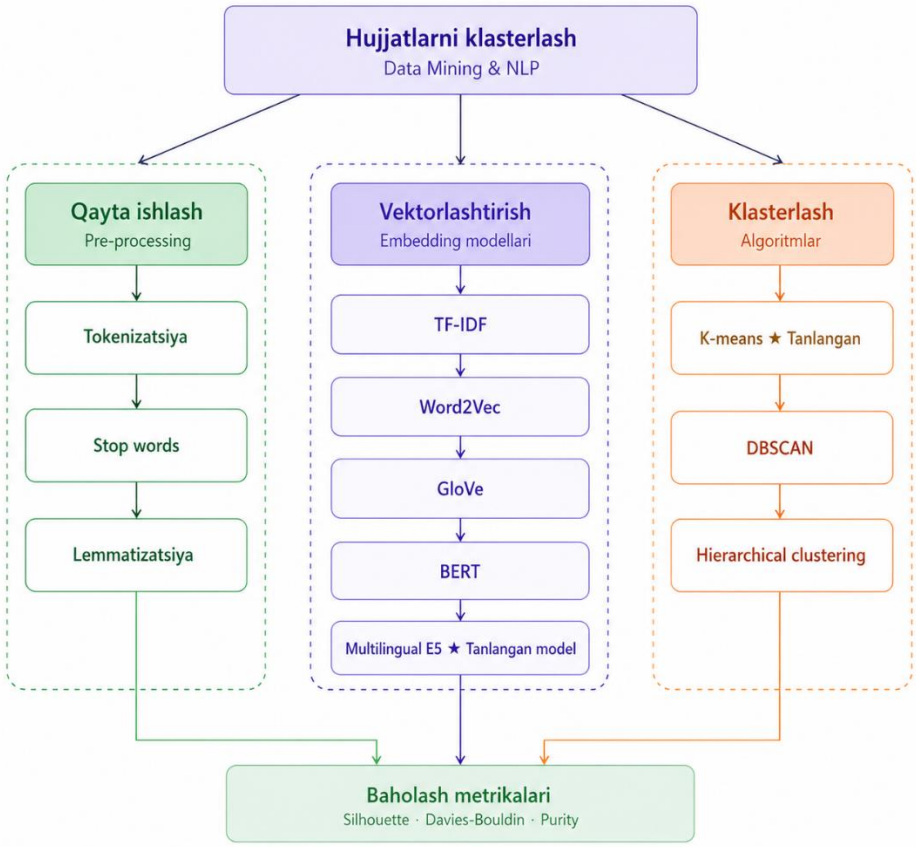
discusses the development of a system to direct incoming documents in the chancellery to the appropriate clusters within the organization.

Keywords: *e-documents, document clustering, text vectorization, clustering algorithms, natural language processing, text similarity, K-means.*

Kirish

Zamonaviy axborot texnologiyalarining jadal rivojlanishi tashkilotlarda elektron hujjatlar hajmini keskin oshirmoqda. Katta hajmdagi hujjatlarni qo‘lda tasniflash ko‘p vaqt talab etishi, inson omili tufayli xatoliklarga olib kelishi va resurslarni behuda sarflashiga sabab bo‘ladi. Shu muammoni hal etish maqsadida elektron hujjatlarni avtomatik klasterlarga ajratish texnologiyalari muhim ahamiyat kasb etmoqda.

Elektron hujjatlarni klasterlarga ajratish jarayoni bir nechta bosqichlarni o‘z ichiga oladi. Birinchi bo‘lib ma’lumotlarni yig‘ish amalga oshiriladi, keyin matnli ma’lumotlar oldindan qayta ishlanadi, tozalanadi va normallashtiriladi. Keyingi bosqichda hujjatlar mashina tomonidan qayta ishlanishi mumkin bo‘lgan vektorli ko‘rinishga o‘tkaziladi (TF-IDF, Word2Vec, GloVe kabi vektorlashtirish usullari). Hujjatlarni klasterlashdan avval ularning o‘zaro o‘xshashlik darajasini aniqlash talab etiladi (Hujjatlarning bir-biriga qanchalik o‘xshashligini Cosine, Jaccard va Euclidean usullari orqali baholash). Olingan vektorlar asosida klasterlash algoritmlari qo‘llanilib, hujjatlar ma’no mazmuniga ko‘ra guruhlarga ajratiladi, yakunda esa natijalarni baholash amalga oshiriladi (masalan, Silhouette ko‘rsatkichi, Davies-Bouldin indeksi, Purity va boshqalar).



1-rasm: Hujjatlarni kalsterlashga ajratish jarayonlari va algoritmlari

1-rasmda ushbu tadqiqotda taklif etilgan avtomatik hujjat klasterlash tizimining jarayonlari va jarayonlarda foydalanish mumkin bo‘lgan algoritmlar, natijalarni baholash metrikalarining umumiy arxitekturasi tasvirlangan. Tizim matnni oldindan qayta ishlashdan (tokenizatsiya, stopWords olib tashlash va lemmatizatsiya) boshlanib, TF-IDF, Word2Vec, GloVe kabi an’anaviy usullardan BERT va Multilingual E5 kabi zamonaviy transformer arxitekturaga asoslangan modellargacha bo‘lgan vektorlashtirish bosqichi orqali davom etadi. K-means, DBSCAN hamda iyerarxik klasterlashtirish algoritmlarini qo‘llash bilan yakunlanadi. Rasmda ko‘rsatilganidek, ushbu tadqiqotda vektorlashtirish uchun Multilingual E5, klasterlashtirish uchun esa K-means algoritmi tanlangan bo‘lib, pipeline oxirida Silhouette, Davies-Bouldin va Purity metrikalari yordamida natijalar sifati baholanadi. Ushbu jarayonlar tashkilotlarda hujjatlar bilan ishlash samaradorligini oshirishga, qidiruv vaqtini qisqartirishga hamda qaror qabul qilish

jarayonlarini tezlashtirishga yordam beradi. Maqolada tashkilotlarda elektron hujjatlarni klasterlash algoritmlari va modellari tahlil qilinib, eng samarali yechim taklif etiladi.

Xorijiy tadqiqotchilar tomonidan hujjatlarni avtomatik klasterlash muammosi ko‘p yillar davomida o‘rganib kelingan. Dastlabki ishlarda hujjatlarni raqamli ko‘rinishga o‘tkazish (vektorlashtirish) uchun asosan statistik usullardan foydalanilgan. Jumladan, *Term Frequency-Inverse Document Frequency* (TF-IDF) usuli bu davrning eng keng tarqalgan yondashuvi hisoblanadi.

Tadqiqotlarga ko‘ra, TF-IDF usuli har bir so‘zga uning hujjatdagi chastotasi va butun korpusdagi kamyobligiga qarab og‘irlik beradi. Ushbu usul hujjat mazmunidagi muhim atamalarni ajratib olishda samarali bo‘lib, K-means algoritmi bilan birgalikda keng qo‘llanilgan [1; 2]. Biroq TF-IDF usulining asosiy kamchiligi – u so‘zlarning semantik ma‘nosini hisobga olmaydi: bir xil ma‘noni anglatuvchi turli so‘zlar (sinonimlar) va turli ma‘noni anglatuvchi bir xil so‘zlar (omonimlar) bir xil tarzda qayta ishlanadi.

2013-yilda Google kompaniyasi tomonidan taqdim etilgan *Word2Vec* modeli so‘zlarni zichlik vektorlari (dense vectors) sifatida ifodalab, semantik munosabatlarni qisman aks ettira oldi. Mikolov va boshqalar tomonidan ishlab chiqilgan ushbu model “qirol − erkak + ayol = malika” kabi matematik munosabatlarni vektoral fazoda namoyish etdi [7; 2]. Word2Vec asosida hujjatlarni raqamli ko‘rinishga o‘tkazish bo‘yicha bir qator xorijiy tadqiqotlar amalga oshirildi. Xususan, Li va boshqalar tomonidan olib borilgan tadqiqotda Word2Vec yordamida olingan so‘z vektorlarini K-means algoritmi bilan birlashtirib, mavzularni ajratib olishda an’anaviy TF-IDF usulidan ustunlik ko‘rsatgani qayd etilgan [5; 1].

Stanford universiteti tomonidan ishlab chiqilgan *GloVe* (Global Vectors for Word Representation) modeli esa so‘zlarning global birgalikda uchrash statistikasiga asoslanib, ularning vektorial tasvirlarini yaratadi [3; 4]. GloVe va

Word2Vec ikkalasi ham “statik embedding” modellari hisoblanadi: bir xil so‘z, kontekstdan qat’iy nazar, doimo bir xil vektorga ega bo‘laveradi. Bu esa ko‘p ma’noli so‘zlarni aniq ifodalashda cheklovlar tug‘diradi. Masalan, K. Sarıtaş va C.A. Öz o‘z tadqiqotlarida turk tilidagi “yaz” so‘zini misol sifatida keltiradi: “Numarayı bir deftere yaz” (“Raqamni daftarga yoz”) jumlasida “yaz” fe’l ma‘nosini “yozmoq” anglatsa, “Bu yaz tatile çıkacağım” (“Bu yoz ta’tilga chiqaman”) jumlasida esa fasl ma’nosini “yoz” anglatadi. Biroq statik modellar ikkala holatda ham ushbu so‘zga bir xil vektorni beradi [10; 2]. Xuddi shunday, ingliz tilidagi “bank” so‘zi “river bank” (daryo qirg‘og‘i) va “bank account” (bank hisobi) iboralarida tamoman boshqa ma’nolarni anglatsa-da, Word2Vec va GloVe kabi statik modellarda u doimo bir xil vektorial tasvir bilan ifodalanadi [7; 9].

2018-yilda Google tomonidan taqdim etilgan *BERT* (Bidirectional Encoder Representations from Transformers) modeli hujjatlarni vektorlashtirish sohasida yangi davr ochdi. BERT ikki tomonlama transformer arxitekturasiga asoslanib, so‘zni jumlaning to‘liq kontekstida, chap va o‘ng tomonlarini hisobga olgan holda vektorlashtiradi. Natijada bir xil so‘z turli kontekstlarda turli vektorlarga ega bo‘ladi, ya‘ni *kontekstual embedding* hosil qilinadi.

2024-yilda nashr etilgan “Text clustering with large language model embeddings” nomli tadqiqotda turli embedding modellari, TF-IDF, BERT, GPT-3.5 va boshqalar klasterlash sifatiga ta‘siri sinab ko‘rildi. Tadqiqot natijalari shuni ko‘rsatdiki, LLM asosidagi embeddinglar an‘anaviy usullarga nisbatan klasterlar sifatini sezilarli darajada oshiradi. Xususan, yirik modellar ma‘lumotlar murakkabligini chuqurroq aks ettirib, klaster aniqligi bo‘yicha yuqori ko‘rsatkichlarga erishadi [3; 1].

Mazkur tadqiqotda K-means algoritmining GPT asosidagi embeddinglar bilan birgalikda ishlatilishi “oddiy va samarali yechim” sifatida tavsiflanadi. Ko‘p modellar qatori ichida K-means algoritmi o‘rtacha reyting bo‘yicha eng yuqori

o‘rinni egalladi [8; 2]. Bu holat K-means algoritmining zamonaviy vektorlashtirishlar bilan muvofiqligini tasdiqlaydi.

Ko‘p tilli hujjatlarni klasterlash muammosi alohida tadqiqot yo‘nalishi sifatida shakllanmoqda. 2025-yilda nashr etilgan maqolada past resurslarga ega tillardagi hujjatlarni (xususan amhar tili misolida) so‘z embeddinglari va ensiklopedik bilimlar yordamida klasterlash usuli taklif etiladi. Spherical K-means algoritmi qo‘llanilgan ushbu tadqiqotda semantik ma’lumotlarni hisobga olish klasterlash sifatini an’anaviy usullarga nisbatan sezilarli darajada oshirishi aniqlangan [11; 10]

Klasterlash algoritmlari ham hujjatlarni avtomatik guruhlash jarayonida muhim o‘rin tutadi. Xorijiy tadqiqotlarda eng ko‘p qo‘llaniladigan algoritmlardan biri K-means hisoblanadi. Ushbu algoritm o‘zining soddaligi va hisoblash tezligi sababli katta hajmdagi ma’lumotlarni klasterlashda keng qo‘llaniladi. K-means algoritmi berilgan ma’lumotlarni oldindan belgilangan klasterlar soniga ajratib, har bir klasterning markazini aniqlash orqali ishlaydi [6].

Shuningdek, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algoritmi ham xorijiy tadqiqotlarda keng o‘rganilgan bo‘lib, u zichlikka asoslangan yondashuv orqali klasterlarni aniqlaydi va shovqin (noisy) ma’lumotlarni ajratib olish imkonini beradi. Ushbu algoritm klasterlar sonini oldindan belgilashni talab qilmasligi bilan ajralib turadi [2].

Bundan tashqari, iyerarxik (Hierarchical) klasterlash algoritmlari ham hujjatlarni tahlil qilishda qo‘llanilib, ma’lumotlar o‘rtasidagi o‘xshashlikni daraxtsimon tuzilma ko‘rinishida aks ettiradi. Xorijiy tadqiqotlarda bu yondashuv kichik hajmdagi ma’lumotlar to‘plamida samarali natija berishi ta’kidlanadi [4; 3].

Yuqorida ko‘rib chiqilgan klasterlash algoritmlari turli yondashuv va xususiyatlarga ega bo‘lib, ularning samaradorligi ma’lumotlar turi va hajmiga bog‘liq holda farqlanadi. Amaliyotda ushbu algoritmlarni mos vektorlash usullari

bilan uyg‘un holda qo‘llash hujjatlarni yanada aniq va mazmunan to‘g‘ri guruhlash imkonini beradi.

Xulosa

Kelgusida ushbu tadqiqotni yanada rivojlantirish maqsadida hujjatlarni klasterlash jarayonida turli klasterlash algoritmlarini (DBSCAN, iyerarxik klasterlash) qo‘llash va ularning samaradorligini taqqoslash rejalashtirilmoqda. Shuningdek, vektorlash bosqichida zamonaviy transformer asosidagi embedding modellarni chuqurroq o‘rganish va ularning klasterlash natijalariga ta’sirini tahlil qilish rejalashtirilmoqda.

Bundan tashqari, real tashkilot muhitida qo‘llash imkoniyatlarini kengaytirish maqsadida tizimni veb-platforma ko‘rinishida ishlab chiqish hamda foydalanuvchi tomonidan yuklangan hujjatlarni avtomatik ravishda tegishli bo‘limlarga ajratish mexanizmini yaratish rejalashtirilmoqda. Ushbu yondashuv hujjatlar bilan ishlash jarayonini avtomatlashtirish, qidiruv vaqtini qisqartirish va boshqaruv samaradorligini oshirishga xizmat qiladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Curiskis S.A., Drake B., Osborn T.R., Kennedy P.J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit // Information Processing & Management. – 2020. – Vol. 57, No. 2. – P. 102034.

2. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD). – 1996. – P. 226–231.

3. Grootendorst M. et al. Text clustering with large language model embeddings // Array. – 2024. – Vol. 24.

4. Johnson S.C. Hierarchical clustering schemes // Psychometrika. – 1967. – Vol. 32, No. 3. – P. 241–254.

5. Li Y., Cai J., Wang J. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding // Journal of Informetrics. – 2018. – Vol. 12, No. 4.

6. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. – 1967. – P. 281–297.

7. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // ICLR Workshop. – 2013.

8. Petukhova A., Matos-Carvalho J.P., Fachada N. Text clustering with large language model embeddings // International Journal of Cognitive Computing in Engineering. – 2025. – Vol. 6. – P. 100–108.

9. Pennington J., Socher R., Manning C.D. GloVe: Global vectors for word representation // Proceedings of EMNLP. – 2014. – P. 1532–1543.

10. Sarıtaş K., Öz C.A., Güngör T. A comprehensive analysis of static word embeddings for Turkish // Expert Systems with Applications. – 2024. – Vol. 252. – Article 124123.

11. Yohannes D., Sinshaw Y.B., Asefa S.H., Assabie Y. Amharic document clustering using semantic information from neural word embedding and encyclopedic knowledge // Scientific African. – 2025. – Vol. 28. – e02657.